CUNY SPS DATA 621 - CTG5 - HW4

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

April 24th, 2019

Contents

1 DATA EXPLORATION

In this assignment we explore, analyze and model a dataset containing 8,161 observations with 25 variables each representing a customer at an auto insurance company. Two of the 25 features are target varibles and 23 are predictors. One of the target variables, `TARGET_FLAG`, is a binary categorical variable where 1 indicates that the customer has been in a car crash and 0 indicates they have not. The other target, `TARGET_AMT`, is a continuous numerical variable representing the payout amount if the customer was in a car accident. Of the remaining 23 predictor variables, 13 are categorical and 10 are numerical.

Using this data, we will compose and evaluate several types of models with the following objectives: - Logistic classification models that aim to predict the probability that a person will crash their car - Multiple linear regression models that aim to predict the amount of money it will cost if the person does crash their car

The intended use case for these models is actuarial in nature: specifically, to calculate insurance rates commensurate with policyholders' (or policy applicants') potential risk levels, based on attributes such as income, age, distance to work, tenure as customers, etc.

[JO: REWORD SO NOT REDUNDANT WITH ABOVE] [JO: DID WE CLEAN UP THE VARIABLE TYPES PER SLACK?] In the training dataset, there are 23 predictors and 2 response variables - one is binary value that indicates whether claim was made and the other is numerical value indicating the cost of claim.

Table 1: Data Dictionary

VARIABLE	DEFINITION	TYPE
TARGET_FLAG	car crash = 1, no car crash = 0	binary categorical response
TARGET_AMT	car crash cost = >0, no car crash = 0	continuous numerical response
AGE	driver's age - very young/old tend to be risky	continuous numerical predictor
BLUEBOOK	$ value of vehicle	continuous numerical predictor
CAR_AGE	age of vehicle	continuous numerical predictor
CAR_TYPE	type of car (6types)	categorical predictor
CAR_USE	usage of car (commercial/private)	binary categorical predictor
CLM_FREQ	number of claims past 5 years	discrete numerical predictor
EDUCATION	max education level (5types)	categorical predictor
HOMEKIDS	number of children at home	discrete numerical predictor
HOME_VAL	$ home value - home owners tend to drive more responsibly	continuous numerical predictor
INCOME	$ income - rich people tend to get into fewer crashes	continuous numerical predictor
JOB	job category (8types, 1missing) - white collar tend to be safer	categorical predictor
KIDSDRIV	number of driving children - teenagers more likely to crash	discrete numerical predictor
MSTATUS	maritial status - married people drive more safely	catogerical predictor
MVR_PTS	number of traffic tickets	continuous numerical predictor
OLDCLAIM	$ total claims in the past 5 years	continuous numerical predictor
PARENT1	single parent	binary categorical predictor
RED_CAR	a red car	binary categorical predictor
REVOKED	license revoked (past 7 years) - more risky driver	binary categorical predictor
SEX	gender - woman may have less crashes than man	binary categorical predictor
TIF	time in force - number of years being customer	continuous numerical predictor
TRAVTIME	distance to work	continuous numerical predictor
URBANCITY	urban/rural	binary categorical predictor
YOJ	years on job - the longer they stay more safe	continuous numerical predictor

[JO: CLARIFY WHAT WE MEAN BY APPROPRIATE DISTRIBUTION - CONSISTENT BETWEEN TEST AND TRAIN][JO: SHOULD WE USE A VISUAL HERE? WHAT'S THE SECOND POINT GETTING AT?]

The response variable shows appropriate distribution in the training data. We confirm that for the number of target flags are 0 equals the target amount 0.

1.1 Summary Statistics

[JO: IN THE ENSUING SUMMARY STAT TABLES, SHALL WE REMOVE THE SCIENTIFIC NOTATION AND ROUND TO DECIMAL FOR READABILITY?][JO: WHY ARE VARIABLES SPLIT BETWEEN T2.1 AND T2.2? DO THESE PERTAIN TO THE LOGISTIC AND LINEAR MODELS, RESPECTIVELY?]

Table 2: (#tab:t2.1)Summary statistics

	n	min	mean	median	max	sd
TARGET_AMT	8161	0	1504.3	0	107586	4704.0
AGE	8155	16	44.8	45	81	8.6
YOJ	7707	0	10.5	11	23	4.1
INCOME	7716	0	61898.1	54028	367030	47572.7
HOME_VAL	7697	0	154867.3	161160	885282	129123.8
TRAVTIME	8161	5	33.5	33	142	15.9
BLUEBOOK	8161	1500	15709.9	14440	69740	8419.7
TIF	8161	1	5.3	4	25	4.2
OLDCLAIM	8161	0	4037.1	0	57037	8777.1
MVR_PTS	8161	0	1.7	1	13	2.1
CAR_AGE	7651	0	8.3	8	28	5.7

[JO: THINK THIS WOULD BE BETTER AS A SET OF SMALL MULTIPLE HISTOGRAM TABLES][JO: MSTATUS = z_F needs to be cleaned to F]

Table 3: (#tab:t2.2)Summary statistics for Categorical Variables

	EDUCATION	JOB	CAR_TYPE	KIDSDRIV	HOMEKIDS	CLM_FREQ
	<High School :1203	z_Blue Collar:1825	Minivan :2145	0:7180	0:5289	0:5009
	Bachelors :2242	Clerical :1271	Panel Truck: 676	1: 636	1: 902	1: 997
	Masters :1658	Professional :1117	Pickup :1389	2: 279	2:1118	2:1171
	PhD : 728	Manager : 988	Sports Car : 907	3: 62	3: 674	3: 776
	z_High School:2330	Lawyer : 835	Van : 750	4: 4	4: 164	4: 190
	NA	Student : 712	SUV :2294	NA	5: 14	5: 18
	NA	(Other) :1413	NA	NA	NA	NA

Table 4: (#tab:t2.2)Summary statistics for Binary Categorical Variables

	PARENT1	SEX	MSTATUS	CAR_USE	RED_CAR	REVOKED	URBANICITY
	No :7084	M:3786	Yes:4894	Commercial:3029	no :5783	No :7161	Urban:6492
	Yes:1077	F:4375	No :3267	Private :5132	yes:2378	Yes:1000	Rural:1669

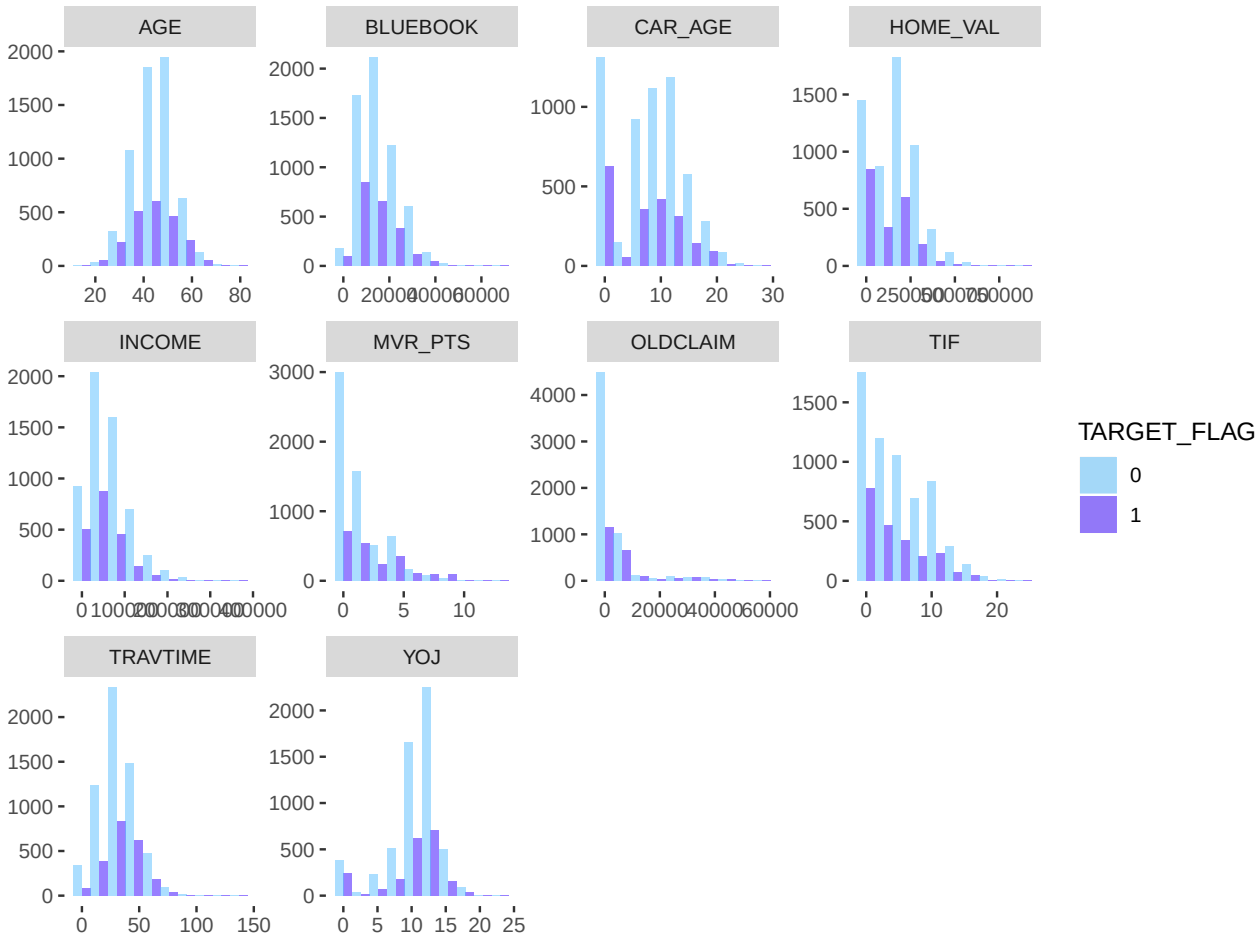


Figure 1: Numeric Data Distributions as a Function of TARGET_FLAG

[JO: IN TERMS OF HIGHER LIKELIHOODS OF ACCIDENT, LOOKS LIKE COMMERCIAL DRIVERS, BLUE COLLAR, UNMARRIED (BECAUSE YOUNGER?), PARENT, REVOKED, MALE, AND URBAN. IS IT WORTH LOOKING AT CONFUSION MATRICES?]

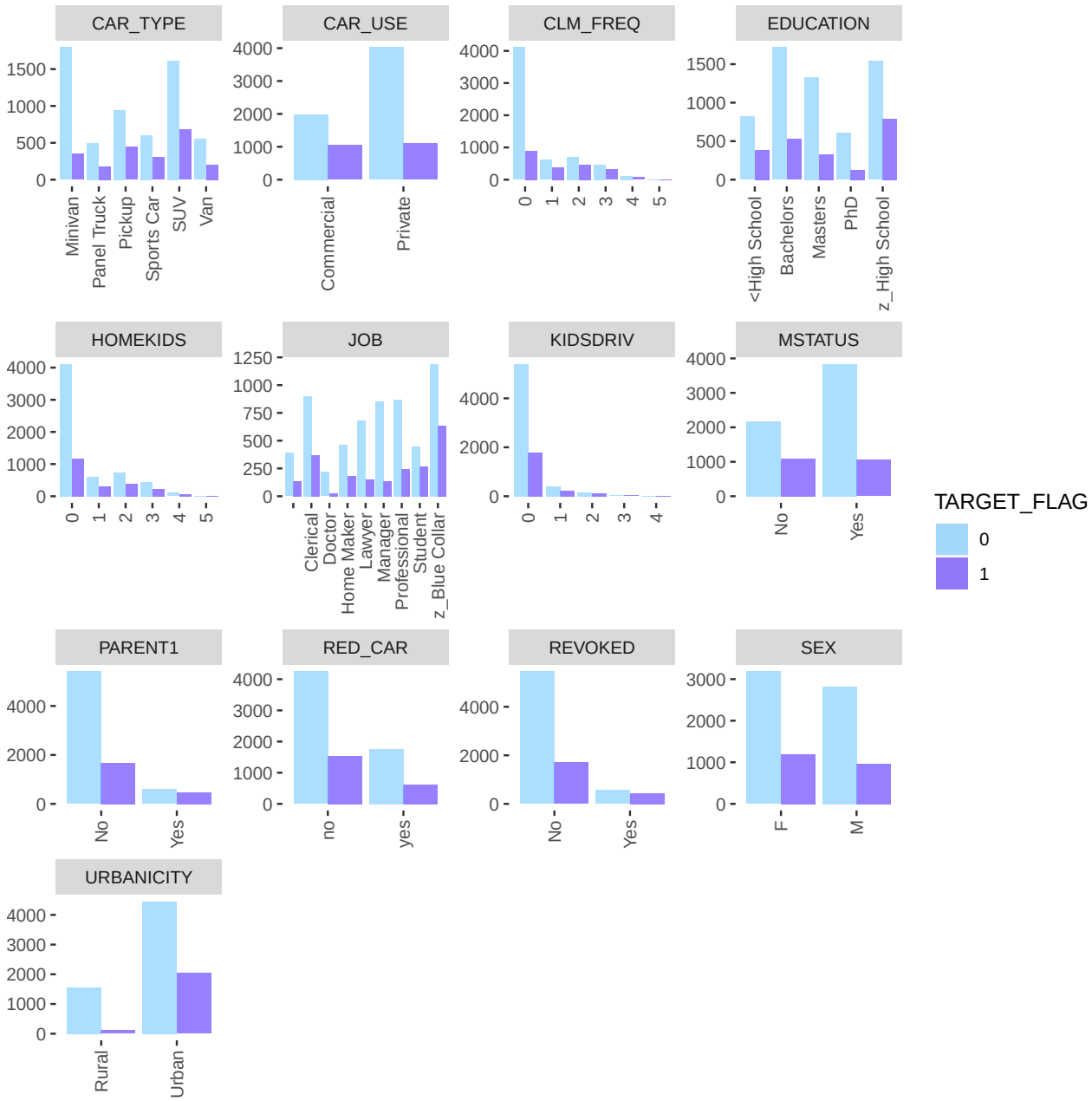


Figure 2: Categorical Data Distributions as a Function of TARGET_FLAG

[JO: IS THIS LIST BASED ONLY ON CONTINUOUS VARIABLES? PERHAPS WE SHOULD BUILD DIFFERENT GRAPHS BASED ON SCALE - ONE FOR YEARS (CAR_AGE, TIF, YOJ, AGE), DOLLARS (OLDCLAIM, BLUEBOOK, INCOME, HOMEVALUE), AND OTHER (MVR_PTS, TRAVTIME - MAYBE SEPARATE?)][JO: WHY DO BLUEBOOK AND INCOME SHOW NO DISTRO?] [JO: WHY INCLUD TARGET_AMT HERE?]

[JO: DO WE NEED THE SCALED VERSION BELOW IF WE SPLIT UP AS ABOVE? WHICH DO WE PREFER?]

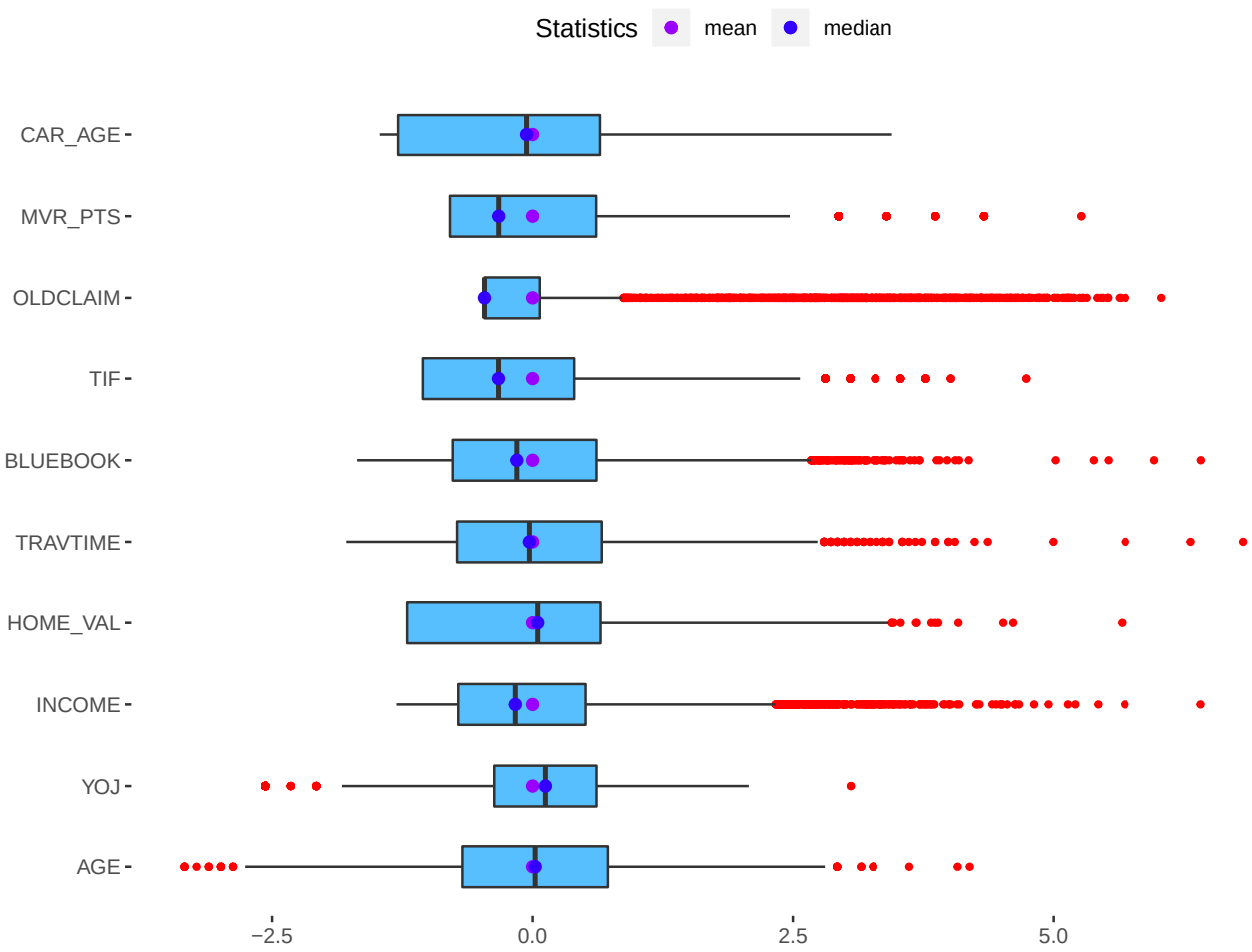


Figure 3: Scaled Boxplots

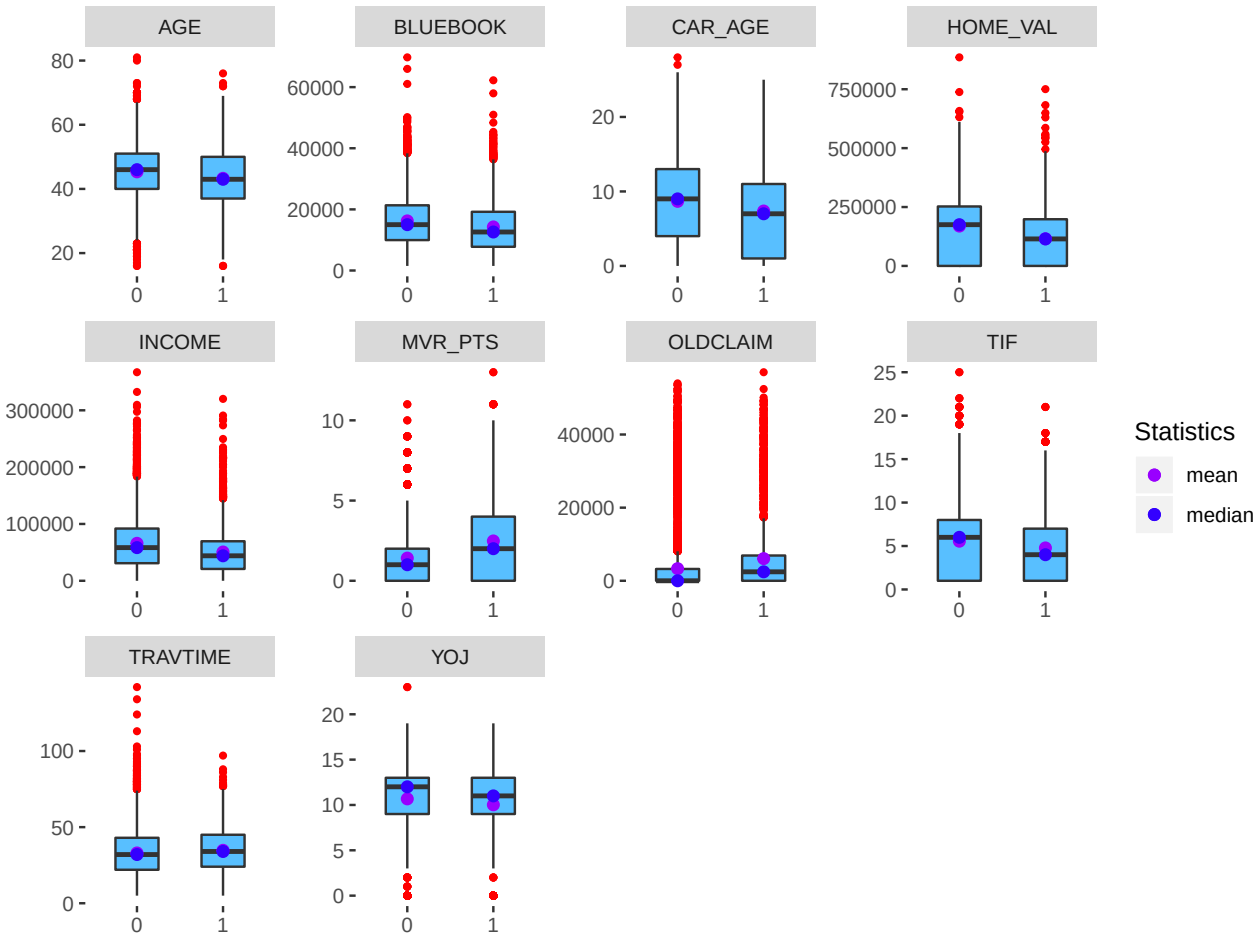


Figure 4: Linear relationship between each numeric predictor and the target

1.2 Linearity

[JO: DUE TO Y-AXIS, HARD TO TELL SLOPE - PERHAPS WE SHOULD ADD SLOPE TO THESE CHARTS TO BETTER DISCERN WHERE THERE SEEMS TO BE A LINEAR RELATIONSHIP?]

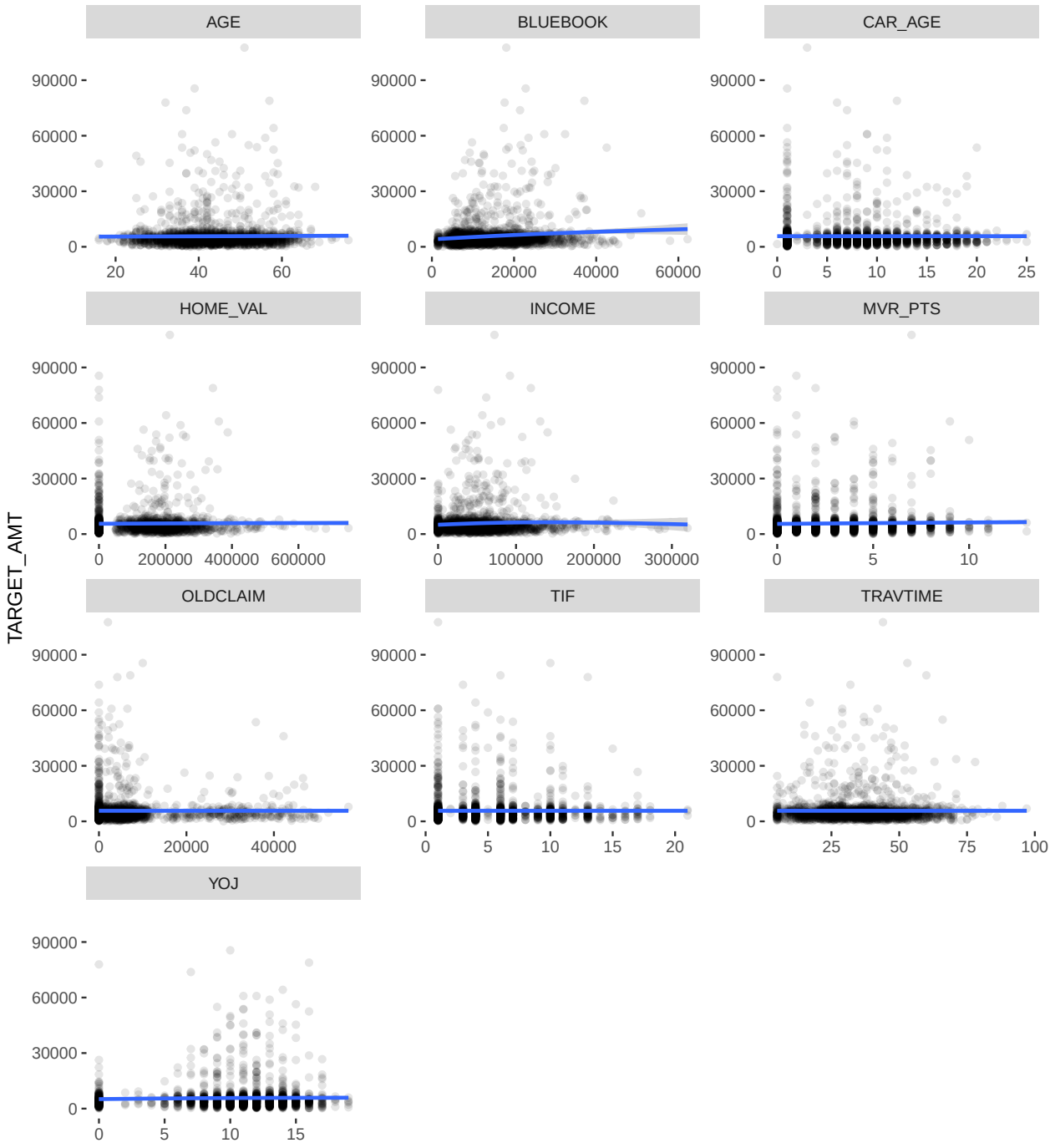


Figure 5: Scatter plot between numeric predictors and the TARGET_AMT

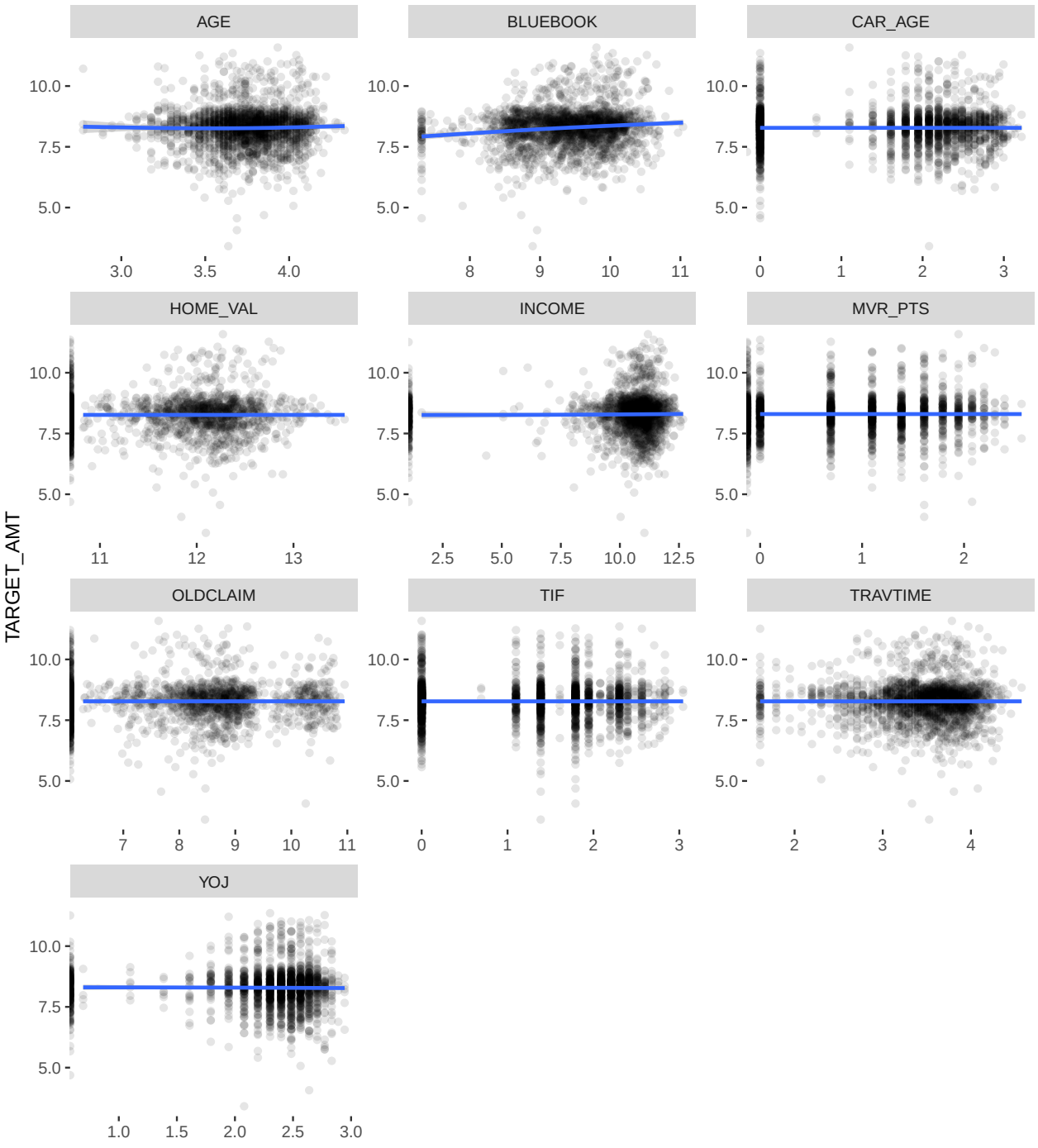


Figure 6: Scatter plot between log transformed numeric predictors and the log transformed TARGET_AMT

1.3 Missing Data

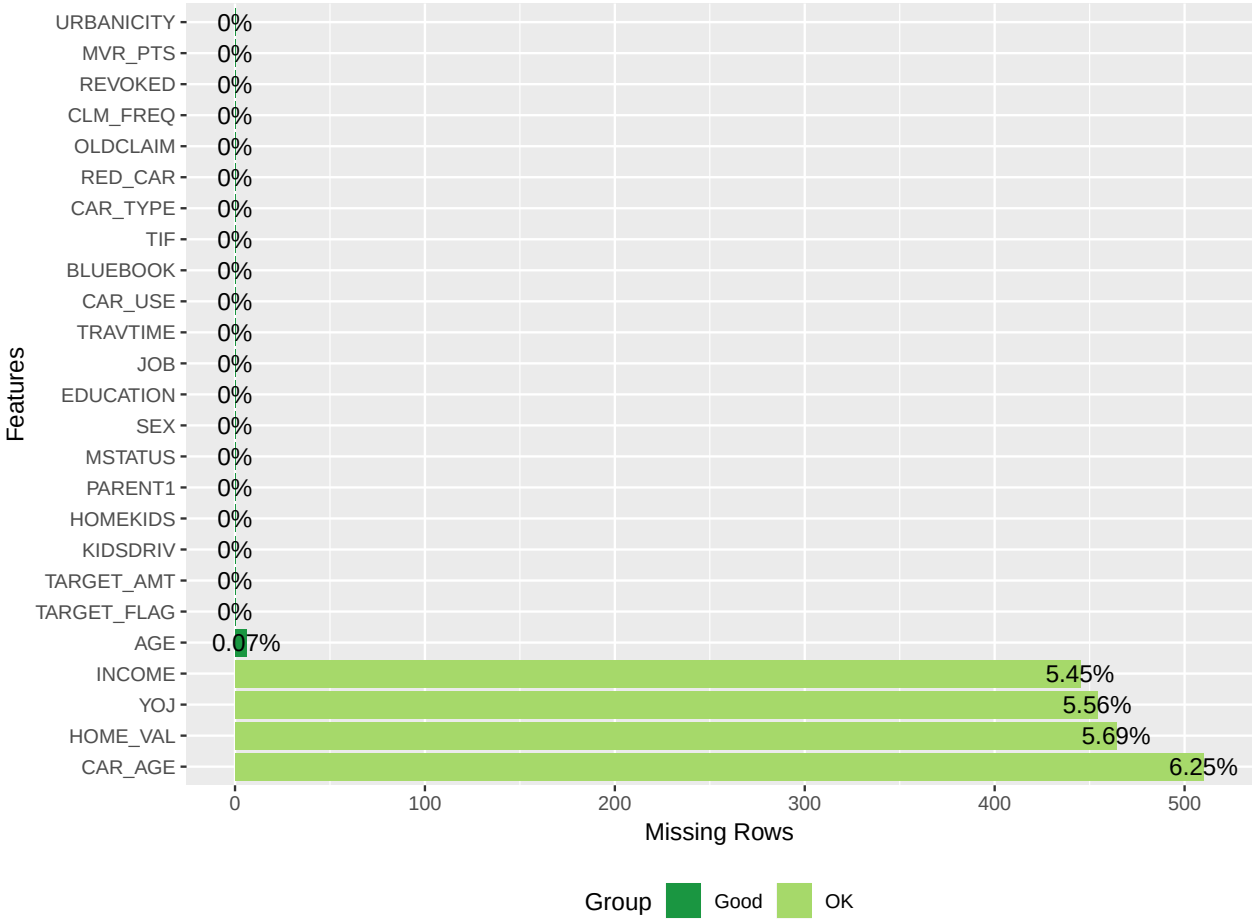


Figure 7: Missing data

There are missing observations for a number of variables: `AGE`, `INCOME`, `YOJ`, `HOME_VAL`, `CAR_AGE`.

[JO: SHALL WE INCLUDE A TABLE THAT SPELLS OUT THE MISSING PROPORTION FOR EACH VARIABLE?][JO: SUGGEST CHECKING THE OVERLAP BETWEEN MISSING VARIABLES I.E. ARE WE MISSING VARIABLES FOR THE SAME OBSERVATIONS? ADDITIONALLY, FOR THOSE MISSING VARIABLES, IS THERE SKEW OR CORRELATION IN OTHER VARIABLES I.E. THEY'RE ALL MARRIED, HIGH EARNERS, HOMEOWNERS?]

[JO: ALIGNED ON DIGGING INTO ABOVE BEFORE MAKING QUALIFICATION BELOW?] Given the low proportion, it seems acceptable to impute the missing values.

2 DATA PREPARATION

2.1 Variable Descriptions

2.1.0.1 KIDSDRIV

`KIDSDRIV` is a categorical predictor with values ranging from 0 to 4. It shows heavy skewness with most cars having 0 kid drivers. Judging from the distribution, it appears that having kid driver results in higher probability of making a claim.

2.1.0.2 AGE

`AGE` presents driver's age and shows normal distribution, centered around 45. Looking at the boxplot of age, there is no difference between the claim made or not in distribution. Therefore, we can believe that `AGE` may not be helpful in determining the probability of making a claim.

2.1.0.3 HOMEKIDS

`HOMEKIDS` is a predictor describing number of children at home ranging from 0 to 5.

2.1.0.4 YOJ

`YOJ` is a predictor describing years on job. It is believed that people who stay at a job for a long time are usually more safe. YOJ shows normal distribution apart from those who are unemployed.

2.1.0.5 INCOME

`INCOME` is a heavily skewed predictor variable. The outliers should be treated.

2.1.0.6 HOME_VAL

`HOME_VAL` is a home value predictor variable. In theory, home owners tend to drive more responsibly. In the graph, we can see difference between the owners and renters.

2.1.0.7 TRAVTIME

`TRAVTIME` is a predictor variable describing the distance to work. Long drives to work usually suggest greater risk. However the graph shows fairly normal distribution and it may not be helpful determining the probability of making a claim.

2.1.0.8 BLUEBOOK

`BLUEBOOK` is a predictor variable describing the value of the car. The‘ boxplot shows that the lower value of the car, the higher chances of making a claim. It is a possibility that the higher price cars are driven more carefully.

2.1.0.9 TIF

`TIF` describes how long the customer has been with the company, and the longer they have, the safer it may be. The plots show the safe drivers tend to stay safe.

2.1.0.10 OLDCLAIM

`OLDCLAIM` is a predictor describing the claims cost made in the past 5 years. We can see that it is very heavily skewed and that most people do not make claims.

2.1.0.11 CLM_FREQ

`CLM_FREQ` is a predictor that describes claim costs in the past 5 years. It seems that people who have made a claim in the past 5 years are highly likely to make another claim.

2.1.0.12 MVR_PTS

`MVR_PTS` is a predictor that describes motor vehicle record points. If you get lots of traffic tickets, you tend to get into more crash. It appears to be a highly significant variable as seen in boxplots.

2.1.0.13 CAR_AGE

`CAR_AGE` describes the vehicle age. There is one data point that shows the vehicle age is -3, this will be corrected to 0.

2.1.0.14 PARENT1

`PARENT1` describes single parent. This is factorized and renamed as `NumParents` to describe the number of parents.

2.1.0.15 SEX

`SEX` describes the gender of the driver. This is factorized and renamed as `MALE` to describe male as 1 and female as 0. It does not appear to be significant variable in the box plot.

2.1.0.16 MSTATUS

`MSTATUS` describes the martial status of the driver. It is believed that married people drive more safely. This variable has been factorized and renamed as `Single` to explain married as 0, not married as 1.

2.1.0.17 EDUCATION

`EDUCATION` describes the education level of the driver. It is factorized. It may be correlated with INCOME.

2.1.0.18 JOB

`JOB` describes the type of job the driver has. It is factorized. It may be correlated with INCOME. In theory white collar jobs tend to drive safer.

2.1.0.19 CAR_TYPE

`CAR_TYPE` describes type of car. It is factorized.

2.1.0.20 CAR_USE

`CAR_USE` describes how the car is used. Commercial vehicles are driven more and may increase probability of collision. It is factorized and renamed as `Commercial`. 0 means private.

2.1.0.21 RED_CAR

`RED_CAR` describes the color of the car is red. It is believed that red cars, especially sports cars are riskier. It is factorized.

2.1.0.22 REVOKED

`REVOKED` describes whether the license has revoked in the past 7 years. If it has revoked, it shows you are a risky driver. It is factorized. The boxplot shows the drivers who had lost their license are likely to be in accidents.

2.1.0.23 URBANICITY

`URBANICITY` describes whether driver lives in Urban area or Rural area. It is factorized and renamed as `URBAN`. 0 means rural.

[JO: THINK WE SHOULD DESCRIBE PURPOSE OF/ NEED FOR VALUE IMPUTATION. MICE IMPUTATION ASSUMES 'MISSING AT RANDOM' (MAR), SO THINK WE'LL NEED TO ESTABLISH THAT THIS IS THE CASE.][JO: WHAT'S THE DIFFERENCE BETWEEN THE M AND MAXIT VALUES (1 FOR AGE, 2 FOR OTHERS?)]

2.2 Missing values

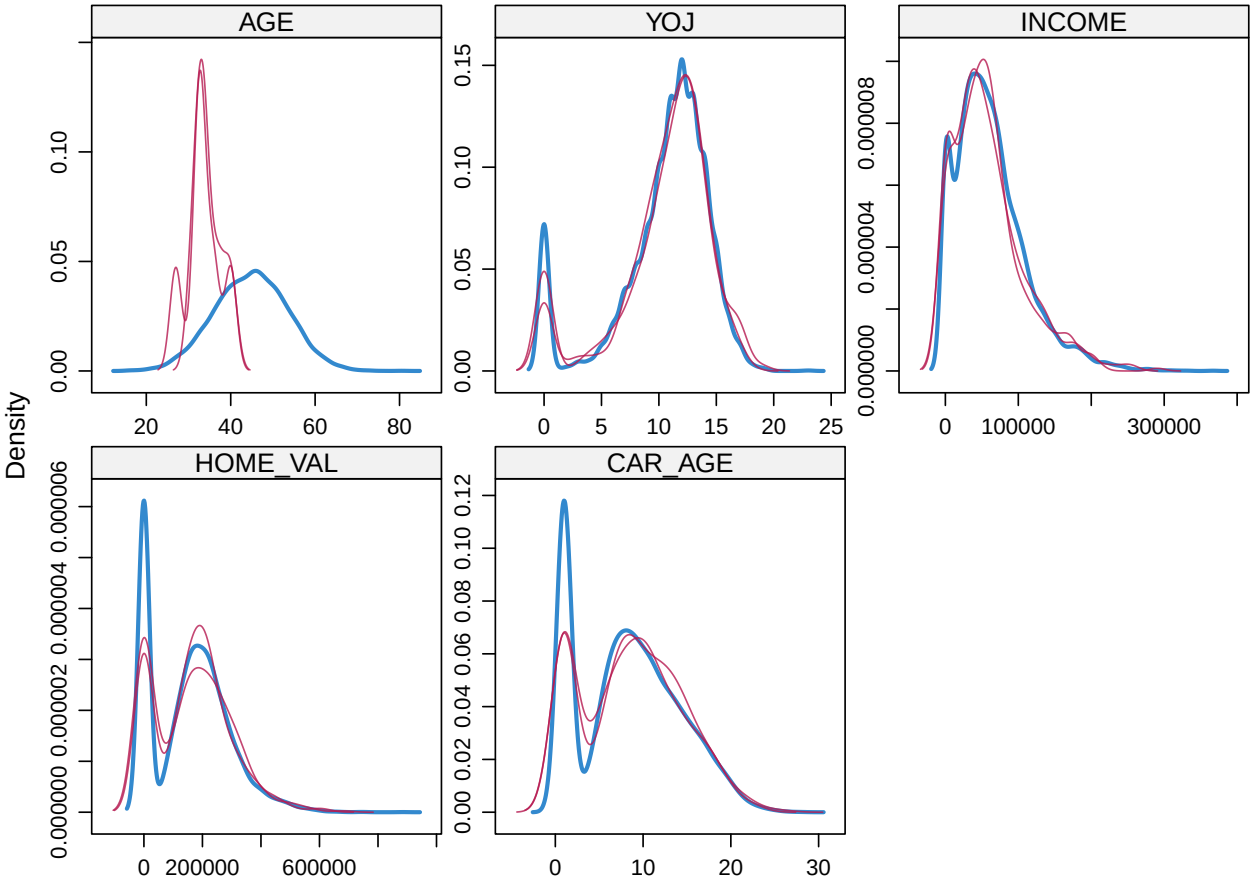
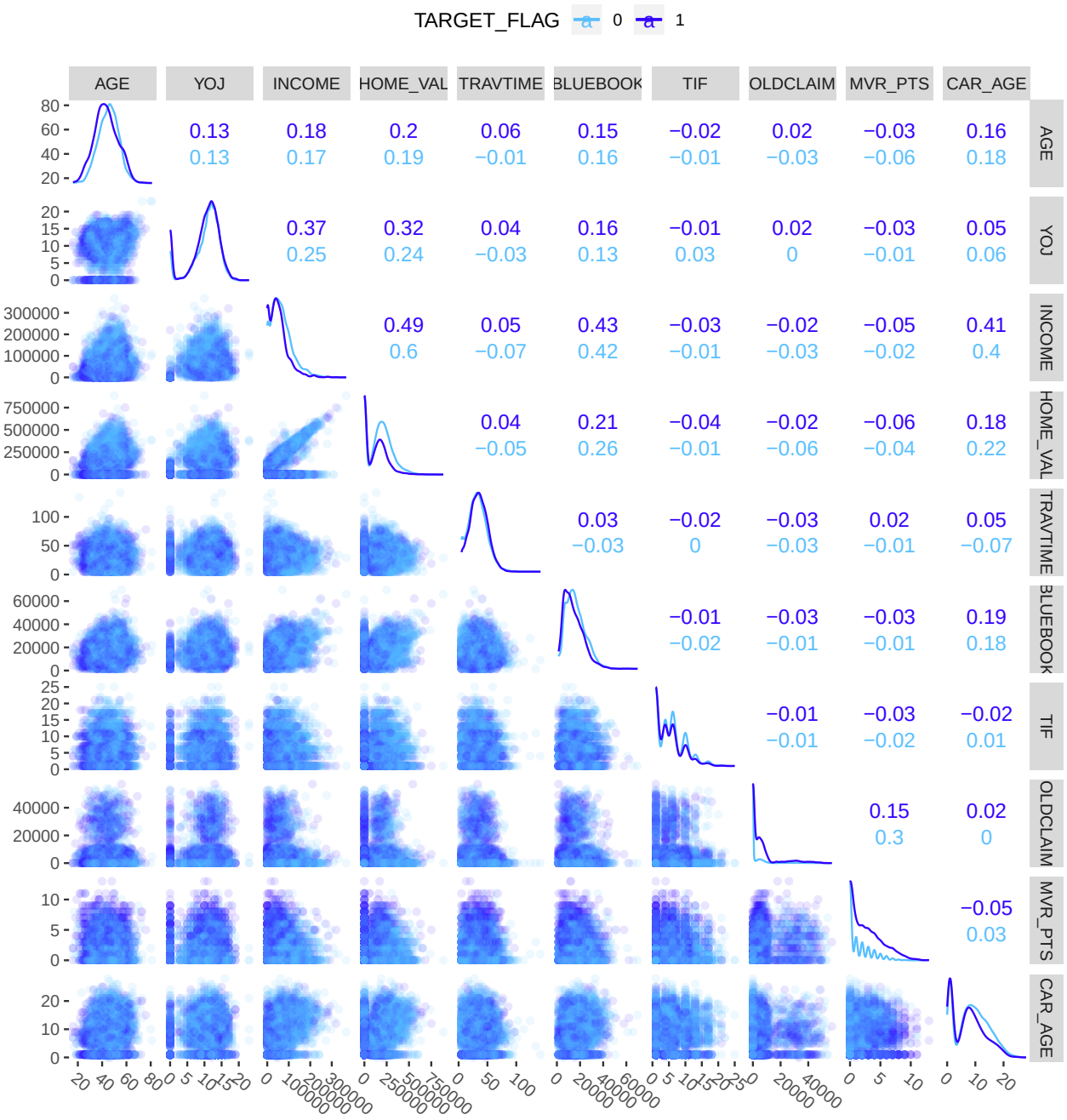


Figure 8: Difference between original and imputed data

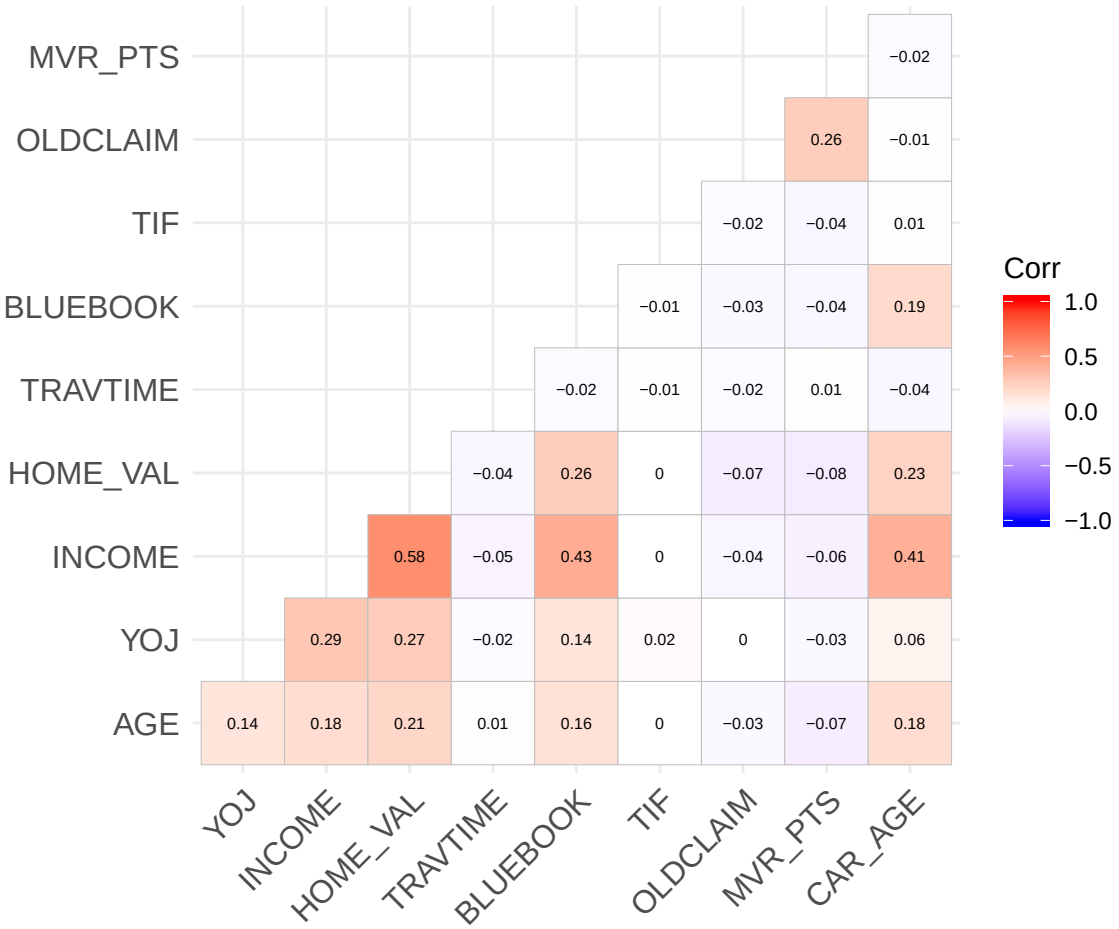
We can see that excpet the `AGE`, the 4 variables roughly matches the existing distribution. We will use the 4 variables and impute `AGE` separately, using the median imputation.

[JO: THINK THE CORR TABLE GETS MUDDLED HERE DUE TO THE LARGE NUMBER OF VARIABLES - WOULD WE CONSIDER JUST RUNNING A CORRPLOT?]



[JO: SHOULD WE THROW THIS CODE BACK IN THE SCRIPT FILE?][JO: WHY THE SUBSET OF VARIABLES HERE?]

Not surpisingly, higher levels of INCOME comes with YOJ; this also means more is disposable, which shows correlation with HOME_VAL and BLUEBOOK. [JO: HOW DOES CAR_AGE CORRELATE WITHINCOME? IS THIS DUE TO HIGHER-END VEHICLES LIVING LONGER? SEEMS SOMEWHAT COUNTERINTUITIVE] Also, MVR_PTS shows relationship with OLDCLAIMS.


MVR_PTS
OLDCLAIM
TIF
BLUEBOOK
TRAVTIME
HOME_VAL
INCOME
YOJ
AGE
YOJ
INCOME
HOME_VAL
TRAVTIME
BLUEBOOK
TIF
OLDCLAIM
MVR_PTS
CAR_AGE
Corr
1.0
0.5
0.0
−0.5
−1.0

3 BUILD MODELS

3.1 Model 1

___ TARGET_FLAG ~ NumParents+ Male+ EDUCATION+ JOB+ CAR_TYPE+ RED_CAR+ REVOKED+ Urban+ Single+ Commercial ___

Model 1 only includes categorical variables as this will be easily interpretable and comprehensible when measuring the leading customers.

[JO: HOW SHALL WE INTERPRET THESE P-VALS - SEEM OUT OF WHACK]

```
##
## Call:
## NULL
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.00234   0.00000   0.00000   0.00000   0.00397
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  398.505  33392.793    0.01     0.99
## TARGET_AMT                  1517.451  12384.946    0.12     0.90
## KIDSDRIV1                      0.331    529.305    0.00     1.00
## KIDSDRIV2                      0.201    836.229    0.00     1.00
## KIDSDRIV3                      2.806    135.962    0.02     0.98
## KIDSDRIV4                      1.104   2110.623    0.00     1.00
## AGE                           -2.331    316.070   -0.01     0.99
## HOMEKIDS1                     -1.820   1299.451    0.00     1.00
## HOMEKIDS2                     -2.197    552.361    0.00     1.00
## HOMEKIDS3                     -2.849    464.703   -0.01     1.00
## HOMEKIDS4                     -0.167    620.815    0.00     1.00
## HOMEKIDS5                    -31.078   6494.202    0.00     1.00
## YOJ                           -4.140    519.439   -0.01     0.99
## INCOME                        -1.598    808.016    0.00     1.00
## PARENT1Yes                    -3.438    958.371    0.00     1.00
## HOME_VAL                       4.646    660.344    0.01     0.99
## MSTATUSNo                      3.277    247.069    0.01     0.99
## SEXF                          -2.317   1693.290    0.00     1.00
## EDUCATIONBachelors            -3.538    807.435    0.00     1.00
## EDUCATIONMasters              -1.178   2616.745    0.00     1.00
## EDUCATIONPhD                   2.305   1154.642    0.00     1.00
## `EDUCATIONz_High School`       2.848    361.704    0.01     0.99
## JOBClerical                   67.953  18588.446    0.00     1.00
## JOBDoctor                    -12.042  14133.089    0.00     1.00
## `JOBHome Maker`               45.000  13784.107    0.00     1.00
## JOBLawyer                     53.862  15446.201    0.00     1.00
## JOBManager                    58.276  16750.283    0.00     1.00
## JOBProfessional               58.417  17737.410    0.00     1.00
## JOBStudent                    50.814  14475.234    0.00     1.00
## `JOBz_Blue Collar`            79.716  21356.275    0.00     1.00
## TRAVTIME                      -1.795    262.354   -0.01     0.99
## CAR_USEPrivate                -1.544    302.623   -0.01     1.00
## BLUEBOOK                     -17.040    469.637   -0.04     0.97
```

```
## TIF                          0.112    290.203    0.00     1.00
## `CAR_TYPEPanel Truck`        4.313 109683.755    0.00     1.00
## CAR_TYPEPickup               4.005    298.882    0.01     0.99
## `CAR_TYPESports Car`         3.446   1126.379    0.00     1.00
## CAR_TYPEVan                  1.931   1306.329    0.00     1.00
## CAR_TYPESUV                 -2.796   1450.488    0.00     1.00
## RED_CARyes                  -3.173    245.003   -0.01     0.99
## OLDCLAIM                    -3.020    252.822   -0.01     0.99
## CLM_FREQ1                    6.016    304.270    0.02     0.98
## CLM_FREQ2                    4.471    456.965    0.01     0.99
## CLM_FREQ3                    6.357    349.212    0.02     0.99
## CLM_FREQ4                    2.695    375.058    0.01     0.99
## CLM_FREQ5                    1.594   3151.401    0.00     1.00
## REVOKEDYes                   3.305    338.469    0.01     0.99
## MVR_PTS                      0.404    203.323    0.00     1.00
## CAR_AGE                      4.084    437.721    0.01     0.99
## URBANICITYRural             -3.120    289.401   -0.01     0.99
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9417.962292043  on 8160  degrees of freedom
## Residual deviance:    0.000078951  on 8111  degrees of freedom
## AIC: 100
##
## Number of Fisher Scoring iterations: 25
```

[JO: WHAT DO SINGLE TERM DELETIONS REFER TO?]

Df	Deviance	AIC
	7.54e+03	7.61e+03
1	7.54e+03	7.62e+03
1	7.54e+03	7.62e+03
1	7.62e+03	7.7e+03
4	7.58e+03	7.65e+03
8	7.65e+03	7.71e+03
5	7.67e+03	7.74e+03
1	7.61e+03	7.68e+03
1	7.54e+03	7.61e+03
1	7.62e+03	7.7e+03
1	8.11e+03	8.19e+03
4	7.58e+03	7.65e+03
5	7.55e+03	7.62e+03
5	7.67e+03	7.73e+03

[JO: CAN LIDIIA / ROSE (WHOEVER BUILT THE MODEL) ADD SOME NOTES, AND THEN I CAN ELABORATE ON THE APPROACH?]

```
##
## Call:  glm(formula = TARGET_FLAG ~ PARENT1 + SEX + MSTATUS + EDUCATION +
##     JOB + CAR_TYPE + CAR_USE + REVOKED + URBANICITY + KIDSDRIV +
##     HOMEKIDS + CLM_FREQ, family = "binomial", data = train.cat.a)
##
## Coefficients:
##            (Intercept)              PARENT1Yes                     SEXF
##                 -1.677                   0.229                  -0.295
##              MSTATUSNo      EDUCATIONBachelors        EDUCATIONMasters
##                  0.687                  -0.510                  -0.446
##           EDUCATIONPhD  EDUCATIONz_High School             JOBClerical
##                 -0.512                  -0.042                   0.619
##              JOBDoctor           JOBHome Maker                JOBLawyer
##                 -0.342                   0.734                   0.184
##             JOBManager         JOBProfessional               JOBStudent
##                 -0.531                   0.269                   0.768
##       JOBz_Blue Collar     CAR_TYPEPanel Truck          CAR_TYPEPickup
##                  0.444                   0.177                   0.611
##     CAR_TYPESports Car             CAR_TYPEVan             CAR_TYPESUV
##                  1.234                   0.436                   0.960
##         CAR_USEPrivate              REVOKEDYes         URBANICITYRural
##                 -0.754                   0.735                  -2.222
##              KIDSDRIV1               KIDSDRIV2               KIDSDRIV3
##                  0.463                   0.710                   1.048
##              KIDSDRIV4               HOMEKIDS1               HOMEKIDS2
##                  1.411                   0.337                   0.228
##              HOMEKIDS3               HOMEKIDS4               HOMEKIDS5
##                  0.207                   0.041                   0.397
##              CLM_FREQ1               CLM_FREQ2               CLM_FREQ3
##                  0.607                   0.638                   0.652
##              CLM_FREQ4               CLM_FREQ5
##                  0.907                   0.900
##
## Degrees of Freedom: 8160 Total (i.e. Null);  8123 Residual
## Null Deviance:       9420
## Residual Deviance: 7540  AIC: 7610
```

[JO: WHERE DOES THE RED_CAR FINDING COME IN?]

`AIC` suggests that `RED_CAR` to be removed.

	x
TARGET_AMT	1251636957159
KIDSDRIV1	2286133776
KIDSDRIV2	5706116953
KIDSDRIV3	150843349
KIDSDRIV4	36350606916
AGE	815183549
HOMEKIDS1	13778749447
HOMEKIDS2	2489638271
HOMEKIDS3	1762143379
HOMEKIDS4	3144953200
HOMEKIDS5	344145223128
YOJ	2201708634
INCOME	5327577150
PARENT1Yes	7494753096
HOME_VAL	3558198539
MSTATUSNo	498111223
SEXF	23396618352
EDUCATIONBachelors	5319921216
EDUCATIONMasters	55874426794
EDUCATIONPhD	10878899203
‘EDUCATIONz_High School‘	1067568094
JOBClerical	2819527314950
JOBDoctor	1629912656747
‘JOBHome Maker‘	1550412993696
JOBLawyer	1946854606926
JOBManager	2289467370504
JOBProfessional	2567264132097
JOBStudent	1709784341702
‘JOBz_Blue Collar‘	3721698310762
TRAVTIME	561648559
CAR_USEPrivate	747299631
BLUEBOOK	1799762578
TIF	687216294
‘CAR_TYPEPanel Truck‘	98169093788096
CAR_TYPEPickup	728934422
‘CAR_TYPESports Car‘	10352835274
CAR_TYPEVan	13924998003
CAR_TYPESUV	17167950261
RED_CARyes	489815817
OLDCLAIM	521578208
CLM_FREQ1	755454299
CLM_FREQ2	1703948261
CLM_FREQ3	995104409
CLM_FREQ4	1147854382
CLM_FREQ5	81039653305
REVOKEDYes	934819142
MVR_PTS	337337273
CAR_AGE	1563454152
URBANICITYRural	683424406

3.2 Model 2

[JO: CAN LIDIIA / ROSE (WHOEVER BUILT THE MODEL) ADD SOME NOTES, AND THEN I CAN ELABORATE ON THE APPROACH?]

```
##
## Call:
## NULL
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.444  -0.714  -0.389   0.639   3.162
##
## Coefficients:
##                           Estimate Std. Error z value             Pr(>|z|)
## (Intercept)               -1.44318    0.03524  -40.95 < 0.0000000000000002
## KIDSDRIV1                  0.12812    0.03076    4.17   0.0000310873201877
## KIDSDRIV2                  0.12971    0.02974    4.36   0.0000128863117635
## KIDSDRIV3                  0.07638    0.02735    2.79              0.00524
## KIDSDRIV4                  0.02420    0.02607    0.93              0.35309
## AGE                        0.01871    0.03605    0.52              0.60385
## HOMEKIDS1                  0.10475    0.03729    2.81              0.00496
## HOMEKIDS2                  0.07850    0.04034    1.95              0.05167
## HOMEKIDS3                  0.06408    0.03751    1.71              0.08757
## HOMEKIDS4                  0.01671    0.03056    0.55              0.58440
## HOMEKIDS5                  0.02532    0.02738    0.92              0.35519
## YOJ                       -0.06375    0.03471   -1.84              0.06626
## INCOME                    -0.14210    0.05356   -2.65              0.00797
## HOME_VAL                  -0.17701    0.04495   -3.94   0.0000820864209119
## TRAVTIME                   0.23292    0.03005    7.75   0.0000000000000091
## BLUEBOOK                  -0.17296    0.04447   -3.89              0.00010
## TIF                       -0.22935    0.03060   -7.50   0.0000000000000658
## OLDCLAIM                  -0.18139    0.03701   -4.90   0.0000009507042294
## CLM_FREQ1                  0.18886    0.03257    5.80   0.0000000066614733
## CLM_FREQ2                  0.21930    0.03317    6.61   0.0000000000381782
## CLM_FREQ3                  0.18165    0.03130    5.80   0.0000000064828689
## CLM_FREQ4                  0.12242    0.02676    4.57   0.0000047697701355
## CLM_FREQ5                  0.05060    0.02583    1.96              0.05015
## MVR_PTS                    0.21340    0.03024    7.06   0.0000000000017182
## CAR_AGE                   -0.03001    0.04284   -0.70              0.48354
## PARENT1Yes                 0.08155    0.04105    1.99              0.04694
## SEXF                      -0.04743    0.05013   -0.95              0.34406
## EDUCATIONBachelors        -0.16310    0.05209   -3.13              0.00174
## EDUCATIONMasters          -0.08990    0.07281   -1.23              0.21696
## EDUCATIONPhD              -0.03166    0.06186   -0.51              0.60880
## `EDUCATIONz_High School`   0.00769    0.04303    0.18              0.85820
## JOBClerical                0.14607    0.07173    2.04              0.04171
## JOBDoctor                 -0.07615    0.04577   -1.66              0.09614
## `JOBHome Maker`            0.05723    0.05714    1.00              0.31655
## JOBLawyer                  0.02262    0.05158    0.44              0.66103
## JOBManager                -0.18117    0.05609   -3.23              0.00124
## JOBProfessional            0.05649    0.06154    0.92              0.35867
## JOBStudent                 0.05338    0.06112    0.87              0.38250
## `JOBz_Blue Collar`         0.13575    0.07766    1.75              0.08044
```

```
## `CAR_TYPEPanel Truck`     0.14745    0.04473    3.30              0.00098
## CAR_TYPEPickup            0.20708    0.03795    5.46   0.0000000485788736
## `CAR_TYPESports Car`      0.32271    0.04092    7.89   0.0000000000000031
## CAR_TYPEVan               0.17508    0.03667    4.77   0.0000017998034031
## CAR_TYPESUV               0.34250    0.05016    6.83   0.0000000000086400
## REVOKEDYes                0.31479    0.03053   10.31 < 0.0000000000000002
## URBANICITYRural          -0.94650    0.04569  -20.72 < 0.0000000000000002
## MSTATUSNo                 0.26094    0.04359    5.99   0.0000000021484384
## CAR_USEPrivate           -0.36528    0.04447   -8.21 < 0.0000000000000002
##
## (Intercept)              ***
## KIDSDRIV1                ***
## KIDSDRIV2                ***
## KIDSDRIV3                **
## KIDSDRIV4
## AGE
## HOMEKIDS1                **
## HOMEKIDS2                .
## HOMEKIDS3                .
## HOMEKIDS4
## HOMEKIDS5
## YOJ                      .
## INCOME                   **
## HOME_VAL                 ***
## TRAVTIME                 ***
## BLUEBOOK                 ***
## TIF                      ***
## OLDCLAIM                 ***
## CLM_FREQ1                ***
## CLM_FREQ2                ***
## CLM_FREQ3                ***
## CLM_FREQ4                ***
## CLM_FREQ5                .
## MVR_PTS                  ***
## CAR_AGE
## PARENT1Yes               *
## SEXF
## EDUCATIONBachelors       **
## EDUCATIONMasters
## EDUCATIONPhD
## `EDUCATIONz_High School`
## JOBClerical              *
## JOBDoctor                .
## `JOBHome Maker`
## JOBLawyer
## JOBManager               **
## JOBProfessional
## JOBStudent
## `JOBz_Blue Collar`       .
## `CAR_TYPEPanel Truck`    ***
## CAR_TYPEPickup           ***
## `CAR_TYPESports Car`     ***
## CAR_TYPEVan              ***
## CAR_TYPESUV              ***
```

```
## REVOKEDYes               ***
## URBANICITYRural          ***
## MSTATUSNo                ***
## CAR_USEPrivate           ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7262.7  on 8113  degrees of freedom
## AIC: 7359
##
## Number of Fisher Scoring iterations: 5
```

[JO: SAME DEAL, I CAN ELABORATE ON BASED ON SOME NOTES]

[JO: SAME DEAL, I CAN ELABORATE ON BASED ON SOME NOTES]

```
##
## Call:  glm(formula = TARGET_FLAG ~ KIDSDRIV + YOJ + INCOME + HOME_VAL +
##     TRAVTIME + BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ + MVR_PTS +
##     PARENT1 + EDUCATION + JOB + CAR_TYPE + REVOKED + URBANICITY +
##     MSTATUS + CAR_USE, family = "binomial", data = train)
##
## Coefficients:
##            (Intercept)              KIDSDRIV1              KIDSDRIV2
##            -0.98224740             0.58826688             0.79680683
##              KIDSDRIV3              KIDSDRIV4                    YOJ
##             0.95122616             1.35182264            -0.01308327
##                 INCOME               HOME_VAL               TRAVTIME
##            -0.00000298            -0.00000141             0.01454680
##               BLUEBOOK                    TIF               OLDCLAIM
##            -0.00002292            -0.05511462            -0.00002040
##              CLM_FREQ1              CLM_FREQ2              CLM_FREQ3
##             0.57660833             0.62424221             0.61798385
##              CLM_FREQ4              CLM_FREQ5                 MVR_PTS
##             0.80961805             1.08363589             0.10004460
##             PARENT1Yes      EDUCATIONBachelors        EDUCATIONMasters
##             0.43909260            -0.39453920            -0.28665235
##           EDUCATIONPhD  EDUCATIONz_High School             JOBClerical
##            -0.18197416             0.01396118             0.40233737
##              JOBDoctor           JOBHome Maker              JOBLawyer
##            -0.44412884             0.19426731             0.06918825
##             JOBManager         JOBProfessional             JOBStudent
##            -0.56476200             0.15526550             0.20460532
##       JOBz_Blue Collar     CAR_TYPEPanel Truck          CAR_TYPEPickup
##             0.31796348             0.59583007             0.54793664
##     CAR_TYPESports Car             CAR_TYPEVan             CAR_TYPESUV
##             0.96478551             0.63753913             0.70338979
##             REVOKEDYes         URBANICITYRural               MSTATUSNo
##             0.95817734            -2.34835529             0.44636602
##         CAR_USEPrivate
##            -0.75093446
##
```

```
## Degrees of Freedom: 8160 Total (i.e. Null);  8121 Residual
## Null Deviance:       9420
## Residual Deviance: 7270  AIC: 7350
```

[JO: ON WHAT BASIS DOES THIS SUGGEST VARIABLE REMOVAL?]

`AIC` suggest to remove `AGE`, `CAR_AGE` and `Male`

	x
KIDSDRIV1	7.7
KIDSDRIV2	7.2
KIDSDRIV3	6.1
KIDSDRIV4	5.5
AGE	10.6
HOMEKIDS1	11.3
HOMEKIDS2	13.3
HOMEKIDS3	11.5
HOMEKIDS4	7.6
HOMEKIDS5	6.1
YOJ	9.8
INCOME	23.4
HOME_VAL	16.5
TRAVTIME	7.4
BLUEBOOK	16.1
TIF	7.6
OLDCLAIM	11.2
CLM_FREQ1	8.7
CLM_FREQ2	9.0
CLM_FREQ3	8.0
CLM_FREQ4	5.8
CLM_FREQ5	5.5
MVR_PTS	7.5
CAR_AGE	15.0
PARENT1Yes	13.8
SEXF	20.5
EDUCATIONBachelors	22.1
EDUCATIONMasters	43.3
EDUCATIONPhD	31.2
‘EDUCATIONz_High School‘	15.1
JOBClerical	42.0
JOBDoctor	17.1
‘JOBHome Maker‘	26.6
JOBLawyer	21.7
JOBManager	25.7
JOBProfessional	30.9
JOBStudent	30.5
‘JOBz_Blue Collar‘	49.2
‘CAR_TYPEPanel Truck‘	16.3
CAR_TYPEPickup	11.8
‘CAR_TYPESports Car‘	13.7
CAR_TYPEVan	11.0
CAR_TYPESUV	20.5
REVOKEDYes	7.6
URBANICITYRural	17.0
MSTATUSNo	15.5
CAR_USEPrivate	16.1

3.3 Model 3

[JO: SAME DEAL: I CAN ELABORATE ON BASED ON SOME NOTES]

```
##
## Call:
## NULL
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.428  -0.714  -0.390   0.640   3.178
##
## Coefficients:
##                           Estimate Std. Error z value             Pr(>|z|)
## (Intercept)               -1.44289    0.03524  -40.95 < 0.0000000000000002
## KIDSDRIV1                  0.13058    0.03031    4.31  0.00001643754938473
## KIDSDRIV2                  0.13181    0.02946    4.47  0.00000768804374280
## KIDSDRIV3                  0.07750    0.02727    2.84              0.00448
## KIDSDRIV4                  0.02523    0.02584    0.98              0.32889
## HOMEKIDS1                  0.09539    0.03462    2.75              0.00587
## HOMEKIDS2                  0.06959    0.03778    1.84              0.06548
## HOMEKIDS3                  0.05547    0.03525    1.57              0.11556
## HOMEKIDS4                  0.01143    0.02940    0.39              0.69759
## HOMEKIDS5                  0.02386    0.02723    0.88              0.38097
## YOJ                       -0.05941    0.03414   -1.74              0.08179
## INCOME                    -0.14614    0.05348   -2.73              0.00628
## HOME_VAL                  -0.17407    0.04485   -3.88              0.00010
## TRAVTIME                   0.23330    0.03003    7.77  0.00000000000000793
## BLUEBOOK                  -0.18888    0.04001   -4.72  0.00000234388495321
## TIF                       -0.22948    0.03059   -7.50  0.00000000000006250
## OLDCLAIM                  -0.18139    0.03699   -4.90  0.00000094145538813
## CLM_FREQ1                  0.18883    0.03256    5.80  0.00000000667660043
## CLM_FREQ2                  0.21944    0.03316    6.62  0.00000000003664048
## CLM_FREQ3                  0.18214    0.03129    5.82  0.00000000584017897
## CLM_FREQ4                  0.12262    0.02675    4.58  0.00000458130464947
## CLM_FREQ5                  0.05030    0.02587    1.94              0.05191
## MVR_PTS                    0.21281    0.03023    7.04  0.00000000000194384
## PARENT1Yes                 0.08213    0.04102    2.00              0.04529
## EDUCATIONBachelors        -0.17538    0.04885   -3.59              0.00033
## EDUCATIONMasters          -0.11139    0.06529   -1.71              0.08800
## EDUCATIONPhD              -0.04631    0.05769   -0.80              0.42215
## `EDUCATIONz_High School`   0.00517    0.04287    0.12              0.90402
## JOBClerical                0.14373    0.07168    2.01              0.04494
## JOBDoctor                 -0.07390    0.04572   -1.62              0.10602
## `JOBHome Maker`            0.05318    0.05680    0.94              0.34913
## JOBLawyer                  0.02434    0.05148    0.47              0.63641
## JOBManager                -0.18007    0.05603   -3.21              0.00131
## JOBProfessional            0.05646    0.06151    0.92              0.35868
## JOBStudent                 0.05249    0.06107    0.86              0.39006
## `JOBz_Blue Collar`         0.13486    0.07762    1.74              0.08231
## `CAR_TYPEPanel Truck`      0.16244    0.04179    3.89              0.00010
## CAR_TYPEPickup             0.20679    0.03792    5.45  0.00000004921121106
## `CAR_TYPESports Car`       0.30294    0.03386    8.95 < 0.0000000000000002
## CAR_TYPEVan                0.18412    0.03543    5.20  0.00000020367139957
## CAR_TYPESUV                0.31338    0.03881    8.07  0.00000000000000068
## REVOKEDYes                 0.31476    0.03051   10.32 < 0.0000000000000002
## URBANICITYRural           -0.94647    0.04569  -20.72 < 0.0000000000000002
## MSTATUSNo                  0.25943    0.04349    5.96  0.00000000244700328
```

```
## CAR_USEPrivate           -0.36513    0.04443   -8.22 < 0.0000000000000002
##
## (Intercept)              ***
## KIDSDRIV1                ***
## KIDSDRIV2                ***
## KIDSDRIV3                **
## KIDSDRIV4
## HOMEKIDS1                **
## HOMEKIDS2                .
## HOMEKIDS3
## HOMEKIDS4
## HOMEKIDS5
## YOJ                      .
## INCOME                   **
## HOME_VAL                 ***
## TRAVTIME                 ***
## BLUEBOOK                 ***
## TIF                      ***
## OLDCLAIM                 ***
## CLM_FREQ1                ***
## CLM_FREQ2                ***
## CLM_FREQ3                ***
## CLM_FREQ4                ***
## CLM_FREQ5                .
## MVR_PTS                  ***
## PARENT1Yes               *
## EDUCATIONBachelors       ***
## EDUCATIONMasters         .
## EDUCATIONPhD
## `EDUCATIONz_High School`
## JOBClerical              *
## JOBDoctor
## `JOBHome Maker`
## JOBLawyer
## JOBManager               **
## JOBProfessional
## JOBStudent
## `JOBz_Blue Collar`       .
## `CAR_TYPEPanel Truck`    ***
## CAR_TYPEPickup           ***
## `CAR_TYPESports Car`     ***
## CAR_TYPEVan              ***
## CAR_TYPESUV              ***
## REVOKEDYes               ***
## URBANICITYRural          ***
## MSTATUSNo                ***
## CAR_USEPrivate           ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7264.5  on 8116  degrees of freedom
```

```
## AIC: 7354
##
## Number of Fisher Scoring iterations: 5
```

[JO: SAME DEAL: I CAN ELBORATE ON BASED ON SOME NOTES]

[JO: SAME DEAL: I CAN ELABORATE ON BASED ON SOME NOTES]

```
##
## Call:  glm(formula = TARGET_FLAG ~ KIDSDRIV + YOJ + INCOME + HOME_VAL +
##     TRAVTIME + BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ + MVR_PTS +
##     PARENT1 + EDUCATION + JOB + CAR_TYPE + REVOKED + URBANICITY +
##     MSTATUS + CAR_USE, family = "binomial", data = train)
##
## Coefficients:
##            (Intercept)               KIDSDRIV1               KIDSDRIV2
##            -0.98224740              0.58826688              0.79680683
##              KIDSDRIV3               KIDSDRIV4                     YOJ
##             0.95122616              1.35182264             -0.01308327
##                 INCOME                HOME_VAL                TRAVTIME
##            -0.00000298             -0.00000141              0.01454680
##               BLUEBOOK                     TIF                OLDCLAIM
##            -0.00002292             -0.05511462             -0.00002040
##              CLM_FREQ1               CLM_FREQ2               CLM_FREQ3
##             0.57660833              0.62424221              0.61798385
##              CLM_FREQ4               CLM_FREQ5                 MVR_PTS
##             0.80961805              1.08363589              0.10004460
##             PARENT1Yes      EDUCATIONBachelors        EDUCATIONMasters
##             0.43909260             -0.39453920             -0.28665235
##           EDUCATIONPhD  EDUCATIONz_High School             JOBClerical
##            -0.18197416              0.01396118              0.40233737
##              JOBDoctor           JOBHome Maker               JOBLawyer
##            -0.44412884              0.19426731              0.06918825
##             JOBManager         JOBProfessional              JOBStudent
##            -0.56476200              0.15526550              0.20460532
##       JOBz_Blue Collar     CAR_TYPEPanel Truck          CAR_TYPEPickup
##             0.31796348              0.59583007              0.54793664
##     CAR_TYPESports Car             CAR_TYPEVan             CAR_TYPESUV
##             0.96478551              0.63753913              0.70338979
##             REVOKEDYes         URBANICITYRural               MSTATUSNo
##             0.95817734             -2.34835529              0.44636602
##         CAR_USEPrivate
##            -0.75093446
##
## Degrees of Freedom: 8160 Total (i.e. Null);  8121 Residual
## Null Deviance:       9420
## Residual Deviance: 7270  AIC: 7350
```

[JO: SAME DEAL: I CAN ELABORATE ON BASED ON SOME NOTES]

	x
KIDSDRIV1	7.5
KIDSDRIV2	7.1
KIDSDRIV3	6.1
KIDSDRIV4	5.5
HOMEKIDS1	9.8
HOMEKIDS2	11.7
HOMEKIDS3	10.1
HOMEKIDS4	7.0
HOMEKIDS5	6.0
YOJ	9.5
INCOME	23.3
HOME_VAL	16.4
TRAVTIME	7.4
BLUEBOOK	13.1
TIF	7.6
OLDCLAIM	11.2
CLM_FREQ1	8.7
CLM_FREQ2	9.0
CLM_FREQ3	8.0
CLM_FREQ4	5.8
CLM_FREQ5	5.5
MVR_PTS	7.5
PARENT1Yes	13.7
EDUCATIONBachelors	19.5
EDUCATIONMasters	34.8
EDUCATIONPhD	27.2
‘EDUCATIONz_High School‘	15.0
JOBClerical	41.9
JOBDoctor	17.1
‘JOBHome Maker‘	26.3
JOBLawyer	21.6
JOBManager	25.6
JOBProfessional	30.9
JOBStudent	30.4
‘JOBz_Blue Collar‘	49.2
‘CAR_TYPEPanel Truck‘	14.2
CAR_TYPEPickup	11.7
‘CAR_TYPESports Car‘	9.4
CAR_TYPEVan	10.2
CAR_TYPESUV	12.3
REVOKEDYes	7.6
URBANICITYRural	17.0
MSTATUSNo	15.4
CAR_USEPrivate	16.1

3.4 Model 4

The forth model is a binary logistic model including all the explanatory variables plus log transformations of our skewed variables: income, travtime, bluebook, oldclaim and age. We used the backward elimination function to refine our model.

4 SELECT MODELS

[JO: TO FOLLOW]

5 Appendix

The appendix is available as script.R file in `project4_insurance` folder.

https://github.com/betsyrosalen/DATA_621_Business_Analyt_and_Data_Mining

Df	Deviance	AIC
	7.26e+03	7.36e+03
4	7.3e+03	7.39e+03
1	7.26e+03	7.36e+03
5	7.27e+03	7.36e+03
1	7.27e+03	7.36e+03
1	7.27e+03	7.36e+03
1	7.28e+03	7.37e+03
1	7.32e+03	7.42e+03
1	7.28e+03	7.37e+03
1	7.32e+03	7.41e+03
1	7.29e+03	7.38e+03
5	7.33e+03	7.41e+03
1	7.31e+03	7.41e+03
1	7.26e+03	7.36e+03
1	7.27e+03	7.36e+03
1	7.26e+03	7.36e+03
4	7.28e+03	7.37e+03
8	7.32e+03	7.4e+03
5	7.35e+03	7.44e+03
1	7.37e+03	7.46e+03
1	7.87e+03	7.97e+03
1	7.3e+03	7.39e+03
1	7.33e+03	7.43e+03

Df	Deviance	AIC
	7.26e+03	7.35e+03
4	7.3e+03	7.38e+03
5	7.27e+03	7.35e+03
1	7.27e+03	7.36e+03
1	7.27e+03	7.36e+03
1	7.28e+03	7.37e+03
1	7.33e+03	7.41e+03
1	7.29e+03	7.38e+03
1	7.32e+03	7.41e+03
1	7.29e+03	7.38e+03
5	7.33e+03	7.41e+03
1	7.31e+03	7.4e+03
1	7.27e+03	7.36e+03
4	7.29e+03	7.37e+03
8	7.32e+03	7.4e+03
5	7.37e+03	7.45e+03
1	7.37e+03	7.46e+03
1	7.87e+03	7.96e+03
1	7.3e+03	7.39e+03
1	7.33e+03	7.42e+03

Observations	8161
Dependent variable	TARGET_FLAG
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(35)$	2140.47
Pseudo-R^2 (Cragg-Uhler)	0.34
Pseudo-R^2 (McFadden)	0.23
AIC	7349.49
BIC	7601.75

	Est.	S.E.	z val.	p	VIF
(Intercept)	2.04	0.79	2.57	0.01	NA
KIDSDRIV1	0.58	0.11	5.54	0.00	1.15
KIDSDRIV2	0.82	0.15	5.40	0.00	1.15
KIDSDRIV3	0.98	0.30	3.25	0.00	1.15
KIDSDRIV4	1.38	1.12	1.23	0.22	1.15
log(AGE)	-0.31	0.16	-2.00	0.05	1.26
YOJ	0.01	0.01	1.30	0.19	2.37
log(INCOME + 0.00000000000001)	-0.02	0.00	-4.07	0.00	3.23
HOME_VAL	-0.00	0.00	-5.25	0.00	1.75
log(TRAVTIME)	0.41	0.05	7.94	0.00	1.03
log(BLUEBOOK)	-0.32	0.06	-5.87	0.00	1.48
TIF	-0.05	0.01	-7.34	0.00	1.01
log(OLDCLAIM + 0.00000000000001)	0.01	0.00	6.24	0.00	1.26
MVR_PTS	0.10	0.01	7.09	0.00	1.24
PARENT1Yes	0.37	0.10	3.66	0.00	1.64
EDUCATIONBachelors	-0.41	0.11	-3.75	0.00	7.47
EDUCATIONMasters	-0.33	0.16	-2.04	0.04	7.47
EDUCATIONPhD	-0.30	0.19	-1.53	0.13	7.47
EDUCATIONz_High School	0.03	0.09	0.35	0.73	7.47
JOBClerical	0.49	0.19	2.53	0.01	26.60
JOBDoctor	-0.39	0.27	-1.48	0.14	26.60
JOBHome Maker	0.15	0.21	0.69	0.49	26.60
JOBLawyer	0.16	0.17	0.96	0.34	26.60
JOBManager	-0.51	0.17	-2.99	0.00	26.60
JOBProfessional	0.22	0.18	1.26	0.21	26.60
JOBStudent	0.12	0.22	0.55	0.58	26.60
JOBz_Blue Collar	0.39	0.19	2.10	0.04	26.60
CAR_TYPEPanel Truck	0.54	0.14	3.76	0.00	2.33
CAR_TYPEPickup	0.58	0.10	5.80	0.00	2.33
CAR_TYPESports Car	0.96	0.11	8.85	0.00	2.33
CAR_TYPEVan	0.65	0.12	5.32	0.00	2.33
CAR_TYPESUV	0.74	0.09	8.57	0.00	2.33
REVOKEDYes	0.71	0.08	8.82	0.00	1.01
URBANICITYRural	-2.35	0.11	-20.86	0.00	1.14
MSTATUSNo	0.46	0.08	5.62	0.00	1.96
CAR_USEPrivate	-0.75	0.09	-8.18	0.00	2.46

Standard errors: MLE